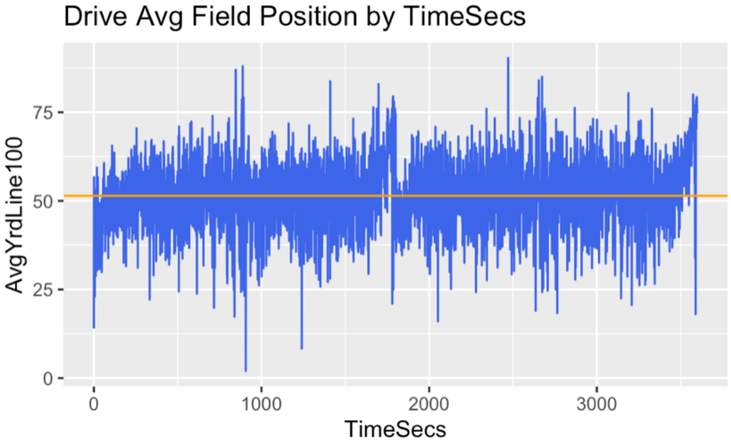
MSDS 456, NFL Assignment, Quinn MacLean

1. Win Probability Model

In creating an NFL Win Probability Model, we will be looking at 2019 play-by-play data using the *nflscrapR* package in R. In cleaning our dataset, we will clean the dataset for any data that doesn't represent an actual play such as Timeouts, Possession change, or No Plays such as penalties. For training our dataset, we will also just consider data within regulation and exclude overtime play. In our exploratory data analysis, we see that **Yards Gained** represents a Poisson Distribution (Appendix A), which could be significant in win probability if there is a "massive gain". It's interesting to see a normal distribution for **ScoreDiff** which goes to show the normalcy and parity of the league. Our preliminary analysis also saw that avg. **ydstogo** and **yrdline100** was relatively even for each team as well as the game progresses, which emphasizes the old adage of a football being a "game of inches".

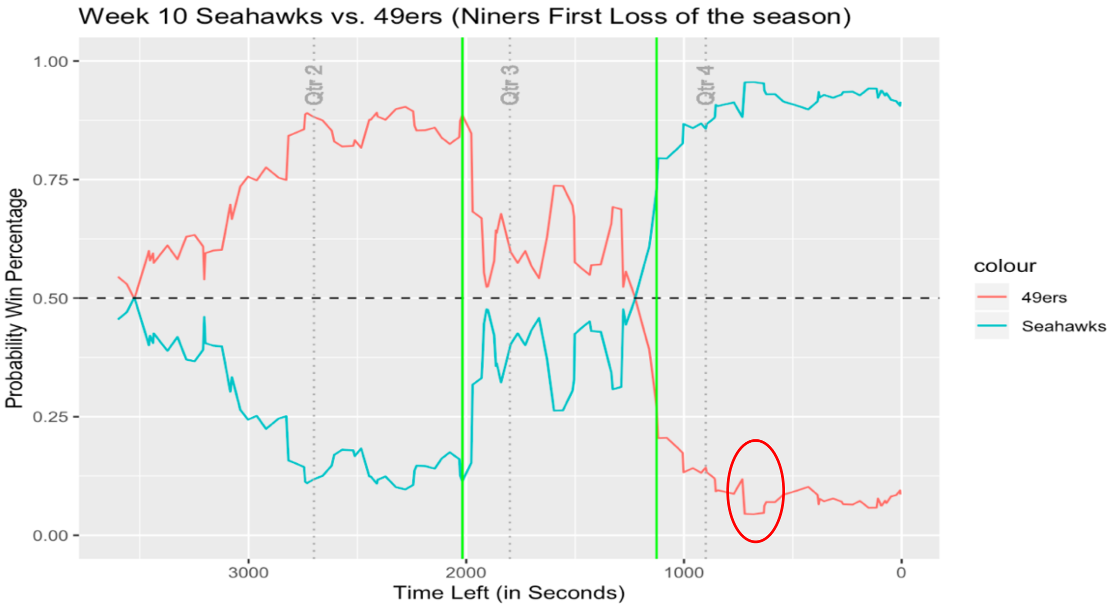


To train the model, we followed a StepAIC GLM approach based on a binomial distribution to predict the winner (binomial due to prediction a win or loss) and to simulate, which variables are statistically significant. From there, we custom selected variables for building the model based on our interpretation of the results of the model (Appendix B). The final model includes down, **TimeSecs** (seconds left), **yrdline100** (position of ball), **ydstogo**, **InterceptionThrown**, **Fumble**, **ScoreDiff**. Both **Yards.Gained** and **ydstogo** had opposite coefficients (**yards.gained** positive, and **ydstogo** negative) and could cause an issue of multicollinearity in the model, so I chose **ytdstogo** as more applicable. Plays that cause turnovers such as **InterceptionThrown** or **Fumble** are big variables for win probability (both negative for possession team). The train set saw a 76.1% accuracy rate, and test set saw a 76.0% accuracy rate for the model, which is a good sign that these percentages are close together as to prove the model isn't overfit due to its simplicity.

2. Win Probability Example

The Week 10 Monday Night game was one of the premier match ups of the season. The San Francisco 49ers lost their first game of the season to division rivals, the Seattle Seahawks. The 49ers controlled the game until the Jadeveon Clowney ran back a 10-yard fumble return for a TD just before half (first green line; **TimeSecs** Left = 2018). That was tipping point for the game as the 49ers had 10-0 lead up until that point. With 4:37 left in 3Q, Russell Wilson found Jacob Hollister for the Seahawks first lead of the game (2nd green line; **TimeSecs** Left = 1126), which shows below given the rise in win probability. The Seahawks added another TD just before the 3Q ended to go up 21-10 heading into the 4th. What's interesting about this probability model is that the 49ers came back. In fact, DeForest Buckner had a fumble return for TD himself (red circle below). The 49ers/Seahawks then traded a few field goals, which is why we still see the

Seahawks still had 93%-win probability heading into overtime (They ended up winning the game in OT; not shown in chart).



3. Model Limitations

One limitation of the model that is talked about in the "Handbook of Statistical Methods and Analyses in Sports" is its simplicity and that any additional variable could have a "multiplicative effect on the size of the state space". There is so many combinations of the variables together that it would difficult to get a sufficient sample size of a combination of events. The book uses for instance of, "how many times has a team been up nine points with a 3rd-and-12 from their own 15 with 3:30 left to play" (Albert, Glickman, Swartz & Koning p.193). This example is to show that specificity is a limitation to the model. One factor I didn't see available in the dataset is who isn't playing. The NFL has a lot of injuries due to its nature of being a physical sport. There is much to be said with who isn't on the field and how that effects win probability then what plays are being conducted to increase the probability.

However, the advantage of the model is for overall situation preparedness. For example, you could run an analysis of 4th down tries and figure out at what yard position is there a point of diminishing returns to win probability. The Oregon Ducks football program have ranked in top 20 (2010 - 2017) for 4th down conversions and it's been a part of the school culture to gamble on 4th down. That's because Chip Kelly realized that if the Ducks increased would increase their scoring differential by going for it, they increase their win probability. Also, he realized that converting a 4th and short scenario was statistically a more likely event then kicking a field to "get the points".

4. Future Analytics Application for the Organization

Chapter 18 of Mathletics dives into team's efficiencies and deficiencies as it relates scoring margin. The concept is simple, win more games by having a better scoring margin (high octane offense, and effective defense is the recipe for success). The regression conducted looks at a team's offense/defense in terms of pass yards/attempt, rushing yards/attempt, turnovers committed, penalty differential, special teams Touchdowns. In creating a regression model, we would be able to quickly see what has attributed to the team's scoring margin. From there it would be necessary to address the team's deficiencies and maintain the team's efficiencies. This is similar to the NBA Four Factors analysis for team construction. It makes sense that the NFL has become more of passive league as the analysis conducted shows that an extra PY/A attempt gain 3x more points than an extra RY/A. A big reason for this could be the shift in defense rules around pass-interference and the no-contact rule, which gives receives a five-yard buffer from the line of scrimmage. Seeing the advantage yardage plays in win probability as earlier, it would make sense to focus on passing efficiency and defensive passing efficiency as an organization to capitalize on this. The last 4 Superbowl's featured elite offensive passing attacks (2019/20 Chiefs, 2018/19 Rams, 2017/18 Eagles, 2016/17 Falcons) and we've seen the NFL become more of a passing attack as a result (2019 passing yds per game avg was 235, 2009 was 218).

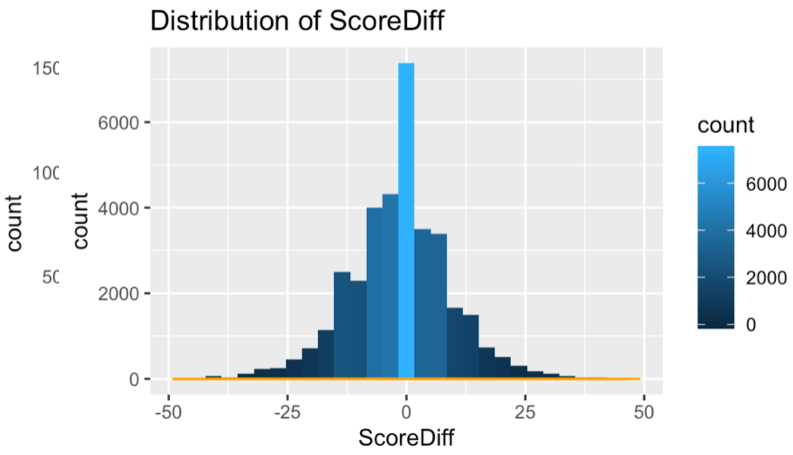
References:

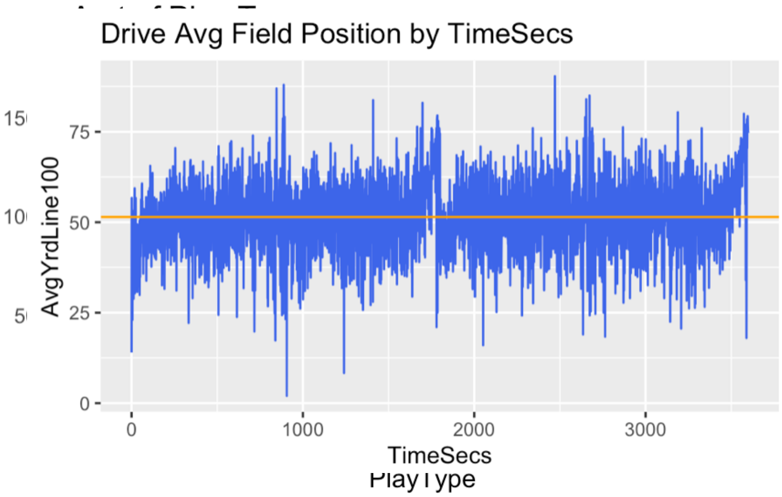
Albert, J. et al. Eds. 2017. Handbook of Statistical Methods and Analyses in Sports. Boca Raton, Fla.: CRC Press. Chapter 8: Situational Success: Evaluating Decision-Making in Football

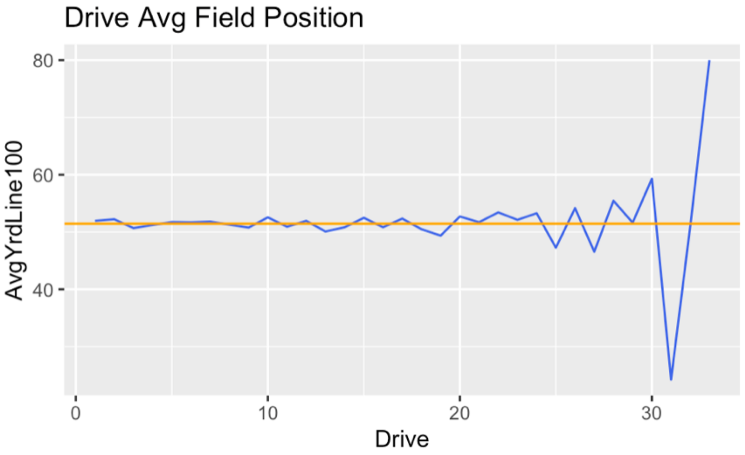
Winston, W. 2012. Mathletics: How Gamblers, Managers, and Sports Enthusiasts Use Mathematics in Baseball, Basketball, and Football. Princeton, NJ.: Princeton Press, P. 127 – 131

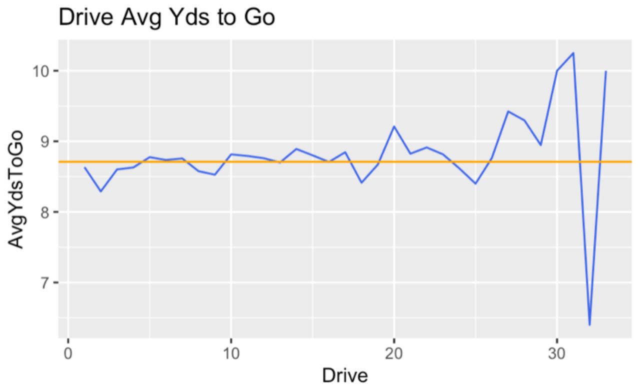
http://www.nfl.com/news/story/09000d5d82a44e69/article/passing-league-explaining-the-nfls-aerial-evolution
https://bleacherreport.com/articles/2840871-ranking-the-top-10-offenses-in-nfl-history#slide5
https://www.pro-football-reference.com/years/NFL/passing.htm
https://247sports.com/college/oregon/ContentGallery/Oregon-Ducks-Football-Breaking-down-a-decade-of-Duck-fourth-downs-120565445/#120565445_1
http://www.thepostgame.com/blog/men-action/201211/how-oregon-coach-chip-kelly-can-spark-moneyball-revolution-nfl

Appendix – Exploratory Data Analysis

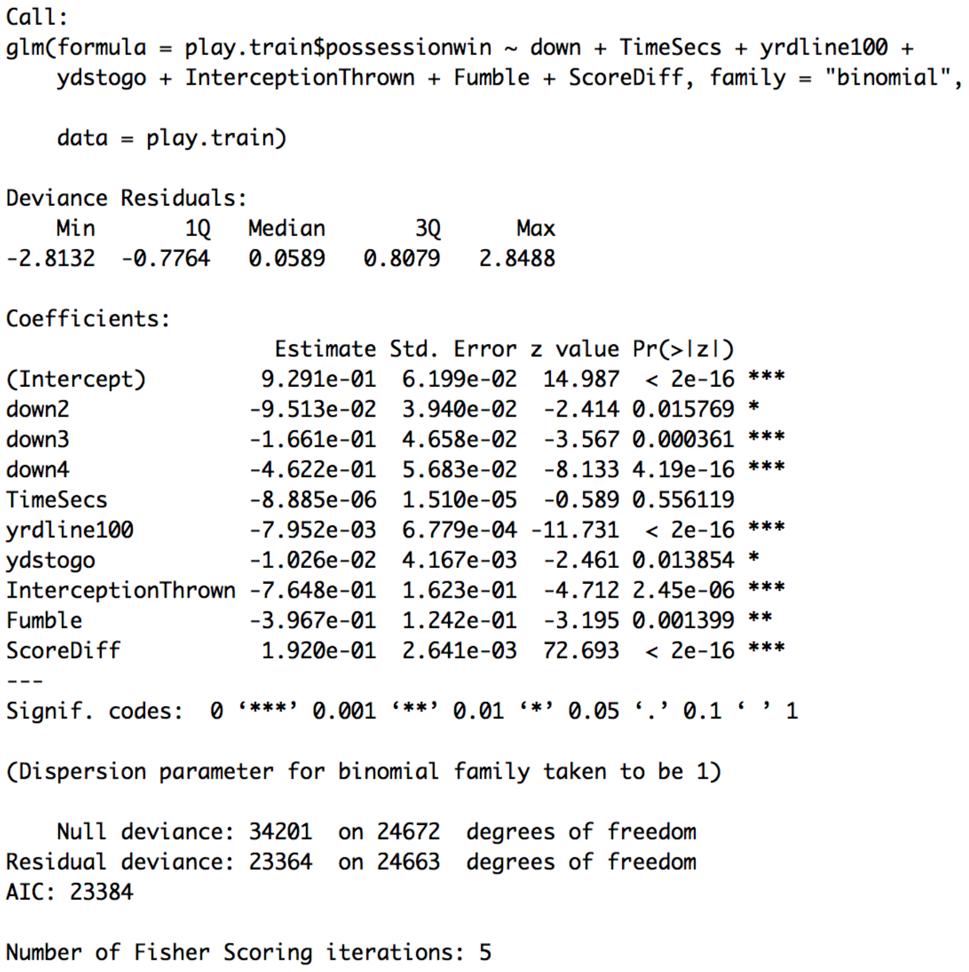

Distribution of ScoreDiff
150
100
50
count
count
6000
4000
2000
0
-50
-25
0
25
50
ScoreDiff
count
6000
4000
2000
0


Drive Avg Field Position by TimeSecs
150
100
50
AvgYrdLine100
75
50
25
0
0
1000
2000
3000
TimeSecs
PlayType


Drive Avg Field Position
AvgYrdLine100
80
60
40
0
10
20
30
Drive


Drive Avg Yds to Go
AvgYdsToGo
10
9
8
7
0
10
20
30
Drive

Appendix B– Final Model Output



```
Call:
glm(formula = play.train$possessionwin ~ down + TimeSecs + yrdline100 +
    ydstogo + InterceptionThrown + Fumble + ScoreDiff, family = "binomial",

    data = play.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8132  -0.7764   0.0589   0.8079   2.8488

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         9.291e-01  6.199e-02  14.987  < 2e-16 ***
down2              -9.513e-02  3.940e-02  -2.414 0.015769 *
down3              -1.661e-01  4.658e-02  -3.567 0.000361 ***
down4              -4.622e-01  5.683e-02  -8.133 4.19e-16 ***
TimeSecs           -8.885e-06  1.510e-05  -0.589 0.556119
yrdline100         -7.952e-03  6.779e-04 -11.731  < 2e-16 ***
ydstogo            -1.026e-02  4.167e-03  -2.461 0.013854 *
InterceptionThrown -7.648e-01  1.623e-01  -4.712 2.45e-06 ***
Fumble             -3.967e-01  1.242e-01  -3.195 0.001399 **
ScoreDiff           1.920e-01  2.641e-03  72.693  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 34201  on 24672  degrees of freedom
Residual deviance: 23364  on 24663  degrees of freedom
AIC: 23384

Number of Fisher Scoring iterations: 5
```